PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 :
C12Q 1/00, 1/34, 1/02, C12P 21/06, 21/04 | A1

(11) International Publication Number: **WO 99/31266**

(43) International Publication Date: 24 June 1999 (24.06.99)

(21) International Application Number: PCT/US98/25862

(22) International Filing Date: 4 December 1998 (04.12.98)

(30) Priority Data:
08/989,380 12 December 1997 (12.12.97) US

(71) Applicant: THE REGENTS OF THE UNIVERSITY OF CALIFORNIA [US/US]; Los Alamos National Laboratory, LC/BPL, MS D412, Los Alamos, NM 87545 (US).

(71)(72) Applicant and Inventor: WALDO, Geoffrey, S. [US/US]; Rt. 1, Box 394C, Espanola, NM 87532 (US).

(74) Agents: FREUND, Samuel, M. et al.; Los Alamos National Laboratory, LC/BPL, MS D412, Los Alamos, NM 87545 (US).

(81) Designated States: AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GE, GH, GM, HR, HU, ID, IL, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, UZ, VN, YU, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published
With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: METHOD FOR DETERMINING AND MODIFYING PROTEIN/PEPTIDE SOLUBILITY

(57) Abstract

A solubility reporter for measuring a protein's solubility *in vivo* or *in vitro* is described. The reporter, which can be used in a single living cell, gives a specific signal which can be used to determine whether the cell bears a soluble version of the protein of interest. A pool of random mutants of an arbitrary protein, generated using error–prone *in vitro* recombination, may also be screened for more soluble versions using the reporter, and these versions may be recombined to yield variants having further enhanced solubility. The method of the present invention includes "irrational" (random mutagenesis) methods, which do not require *a priori* knowledge of the three–dimensional structure of the protein of interest. Multiple sequences of mutation/genetic recombination and selection for improved solubility yield versions of the protein which display enhanced solubility.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	**ES**	Spain	**LS**	Lesotho	**SI**	Slovenia
AM	Armenia	**FI**	Finland	**LT**	Lithuania	**SK**	Slovakia
AT	Austria	**FR**	France	**LU**	Luxembourg	**SN**	Senegal
AU	Australia	**GA**	Gabon	**LV**	Latvia	**SZ**	Swaziland
AZ	Azerbaijan	**GB**	United Kingdom	**MC**	Monaco	**TD**	Chad
BA	Bosnia and Herzegovina	**GE**	Georgia	**MD**	Republic of Moldova	**TG**	Togo
BB	Barbados	**GH**	Ghana	**MG**	Madagascar	**TJ**	Tajikistan
BE	Belgium	**GN**	Guinea	**MK**	The former Yugoslav Republic of Macedonia	**TM**	Turkmenistan
BF	Burkina Faso	**GR**	Greece			**TR**	Turkey
BG	Bulgaria	**HU**	Hungary	**ML**	Mali	**TT**	Trinidad and Tobago
BJ	Benin	**IE**	Ireland	**MN**	Mongolia	**UA**	Ukraine
BR	Brazil	**IL**	Israel	**MR**	Mauritania	**UG**	Uganda
BY	Belarus	**IS**	Iceland	**MW**	Malawi	**US**	United States of America
CA	Canada	**IT**	Italy	**MX**	Mexico	**UZ**	Uzbekistan
CF	Central African Republic	**JP**	Japan	**NE**	Niger	**VN**	Viet Nam
CG	Congo	**KE**	Kenya	**NL**	Netherlands	**YU**	Yugoslavia
CH	Switzerland	**KG**	Kyrgyzstan	**NO**	Norway	**ZW**	Zimbabwe
CI	Côte d'Ivoire	**KP**	Democratic People's Republic of Korea	**NZ**	New Zealand		
CM	Cameroon			**PL**	Poland		
CN	China	**KR**	Republic of Korea	**PT**	Portugal		
CU	Cuba	**KZ**	Kazakstan	**RO**	Romania		
CZ	Czech Republic	**LC**	Saint Lucia	**RU**	Russian Federation		
DE	Germany	**LI**	Liechtenstein	**SD**	Sudan		
DK	Denmark	**LK**	Sri Lanka	**SE**	Sweden		
EE	Estonia	**LR**	Liberia	**SG**	Singapore		

METHOD FOR DETERMINING AND MODIFYING PROTEIN/PEPTIDE SOLUBILITY

FIELD OF THE INVENTION

The present invention relates generally to improving the solubility of proteins/peptides and, more particularly to a method for identifying more or less soluble proteins/peptides from libraries of mutants thereof generated from the directed evolution of genes which express these proteins/peptides. This invention was made with government support under Contract No. W-7405-ENG-36 awarded by the U.S. Department of Energy to The Regents of the University of California. The government has certain rights in the invention.

BACKGROUND OF THE INVENTION

Protein insolubility constitutes a significant problem in basic and applied bioscience, in many situations limiting the rate of progress in these areas. Protein folding and solubility has been the subject of considerable theoretical and empirical research. However, there still exists no general method for improving *intrinsic* protein solubility. Such a method would greatly facilitate protein structure-function studies, drug design, *de novo* peptide and protein design and associated structure-function studies, industrial process optimization using bioreactors and microorganisms, and many disciplines in which a process or application depends on the ability to tailor or improve the solubility of proteins, screen or modify the solubility of large numbers of unique proteins about which little or no structure-function information is available, or adapt the solubility of proteins to new environments when the structure and function of the protein(s) are poorly understood or unknown.

Overexpression of cloned genes using an expression host, for example *E. coli*, is the principal method of obtaining proteins for most applications. Unfortunately, many such cloned foreign proteins are insoluble or unstable when overexpressed. There are two sets of approaches currently in use which deal with such insoluble proteins. One set of approaches modifies the environment of the protein *in vivo* and/or *in vitro*. For example, proteins may be expressed as fusions with more soluble proteins, or directed to specific cellular locations. Chaperons may be coexpressed to assist folding

pathways. Insoluble proteins may be purified from inclusion bodies using denaturants and the protein subsequently refolded in the absence of the denaturant. Modified growth media and/or growth conditions can sometimes improve the folding and solubility of a foreign protein. However, these methods are frequently cumbersome, unreliable, ineffective, or lack generality. A second set of approaches changes the sequence of the expressed protein. Rational approaches employ site-directed mutation of key residues to improve protein stability and solubility. Alternatively, a smaller, more soluble fragment of the protein may be expressed. These approaches require *a priori* knowledge about the structure of the protein, knowledge which is generally unavailable when the protein is insoluble. Furthermore, rational design approaches are best applied when the problem involves only a small number of amino-acid changes. Finally, even when the structure is known, the changes required to improve solubility may be unclear. Thus, many thousands of possible combinations of mutations may have to be investigated leading to what is essentially an "irrational" or random mutagenesis approach. Such an approach requires a method for rapidly determining the solubility of each version.

Random or "irrational" mutagenesis redesign of protein solubility carries the possibility that the native function of the protein may be destroyed or modified by the inadvertent mutation of residues which are important for function, but not necessarily related to solubility. However, protein solubility is strongly influenced by interaction with the environment through surface amino acid residues, while catalytic activities and/or small substrate recognition often involve partially buried or cleft residues distant from the surface residues. Thus, in many situations, rational mutation of proteins has demonstrated that the solubility of a protein can be modified without destroying the native function of the protein. Modification of the function of a protein without effecting its solubility has also been frequently observed. Furthermore, spontaneous mutants of proteins bearing only 1 or 2 point mutations have been serendipitously isolated which have converted a previously insoluble protein into a soluble one. This suggests that the solubility of a protein can be optimized with a low level of mutation and that protein

function can be maintained independently of enhancements or modifications to solubility. Furthermore, a screen for function may be applied concomitantly after each round of solubility selection during the directed evolution process.

In the absence of a screen for function, for example when the function is unknown, the final version of the protein can be backcrossed against the wild type *in vitro* to remove nonessential mutations. This approach has been successfully applied by Stemmer in "Rapid Evolution Of A Protein *In Vitro* By DNA Shuffling," by W.P.C. Stemmer, Nature **370**, 389 (1994), and in "DNA Shuffling By Random Fragmentation And Reassembly: *In Vitro* Recombination For Molecular Evolution," by W.P.C. Stemmer, Proc. Natl. Acad. Sci. **USA 91**, 10747 (1994) to problems in which the function of a protein had been optimized and it was desired to remove nonessential mutations accumulated during directed evolution. The development of highly specialized protein variants by directed, *in vitro* evolution, which exerts unidirectional selection pressure on organisms, is further discussed in: "Searching Sequence Space: Using Recombination To Search More Efficiently And Thoroughly Instead Of Making Bigger Combinatorial Libraries," by Willem P.C. Stemmer, Biotechnology **13**, 549 (1995); in "Directed Evolution: Creating Biocatalysts For The Future," by Frances H. Arnold, Chemical Engineering Science **51**, 5091 (1996); in "Directed Evolution Of A Fucosidase From A Galactosidase By DNA Shuffling And Screening," by Ji-Hu Zhang et al., Proc. Natl. Acad. Sci. **USA 94**, 4504 (1997); in "Functional And Nonfunctional Mutations Distinguished By Random Combination Of Homologous Genes," by Huimin Zhao and Frances H. Arnold, Proc. Natl. Acad. Sci. **USA 94**, 7007 (1997); and in "Strategies For The In Vitro Evolution of Protein Function: Enzyme Evolution By Random Recombination of Improved Sequences", by Jeff Moore et al., J. Mol. Biol. **272**, 336-346 (1997). Therein, efficient strategies for engineering new proteins by multiple generations of random mutagenesis and recombination coupled with screening for improved variants is described. However, there are no teachings concerning the use of directed evolutionary processes to improve solubility of proteins; rather, the mutagenesis was directed to improvement of protein function. It should be mentioned,

however, that in order for the protein to function properly in any environment, it must be correctly folded and, therefore, soluble.

Finally, for structural determination it is often not necessary or even desirable to have a fully functional version of the protein. If the mutational rate is low (ensured by molecular backcrossing), it is likely that the structure of the wild-type and solubility optimized versions of a protein will be similar. As long as the protein is soluble, and a structure can be obtained, it should then be possible to redesign the solubility of the protein using rational methods, if desired.

Green fluorescent protein has become a widely used reporter of gene expression and regulation. DNA shuffling has been used to obtain a mutant having a whole cell fluorescence 45-times greater than the standard, commercially available plasmid GFP. See, e.g., "Improved Green Fluorescent Protein By Molecular Evolution Using DNA Shuffling," by Andreas Crameri et al., Nature Biotechnology **14**, 315 (1996). The screening process optimizes the function of GFP (green fluorescence), and thus uses a functional screen. Although the screening process coincidentally optimizes the solubility of the GFP, in that the GFP is only fluorescent when properly folded, there is no mention of using soluble GFP as a tag to monitor solubility of other proteins; that is, the function of the protein and not its solubility are being modified. In "Wavelength Mutations And Post-translational Auto-oxidation Of Green Fluorescent Protein," by Roger Heim et al., Proc. Natl. Acad. Sci. **USA 91**, 12501 (1994), GFP was mutagenized and screened for variants with altered absorption or emission spectra. The authors mention that in place of proteins labeled with fluorescent tags to detect location and sometimes their conformational changes both *in vitro* and in intact cells, a possible strategy would be to concatenate the gene for the nonfluorescent protein of interest with the gene for a naturally fluorescent protein and express the fusion product. However, the focus of this paper is the extension of the usefulness of GFP by enabling visualization of differential gene expression and protein localization and measurement of protein association by fluorescence resonance energy transfer, by making available two visibly distinct colors. There is no mention of the use of the gene construct for

solubility determinations. The paper further discusses the expression of GFP in *E. coli* under the control of a T7 promoter, and that the bacteria contained inclusion bodies consisting of protein indistinguishable from jellyfish or soluble recombinant protein on denaturing gels, but that this material was completely nonfluorescent, lacked the visible absorbance bands of the chromophore, and did not become fluorescent when solubilized and subjected to protocols that renature GFP, as opposed to the soluble GFP in the bacteria which undergoes correct folding and, therefore, fluoresces.

Chun Wu et al. in "Novel Green Fluorescent Protein (GFP) Baculovirus Expression Vectors," Gene **190**, 157 (1997), describe the construction of Baculovirus expression vectors which contain GFP as a reporter gene. The authors follow the production and purification of a protein of interest by in-frame cloning of the gene that expresses the protein in insect cells with the GFP open reading frame, thereby permitting visualization of the produced GFP-fusion protein using UV light. However, the purified GFP-XylE fusion protein was found to be insoluble after harvest.

In "Application Of A Chimeric Green Protein Fluorescent Protein To Study Protein-Protein Interactions," by N. Garamszegi et al., Biotechniques **23**, 864 (1997), the authors discuss the fusion between GFP and human calmodulin-like protein (CLP) and show that this protein retains fluorescence and the known characteristics of CLP. That is, the GFP portion remains responsible for efficient fluorescent signals with little or no influence on the properties of the fused protein of interest. The authors maintain that the exhibited GFP fluorescence provides information concerning the maintenance of the GFP structural integrity in the chimeric protein, but does not provide information about the integrity of the entire fusion protein and, in particular, does not allow any statements concerning the maintenance of CLP function or integrity. From these statements, it is clear that this paper does not contemplate the use of the GFP as a solubility reporter for the CLP.

Accordingly, it is an object of the present invention to provide a solubility reporter for rapidly identifying soluble forms of proteins.

Another object of the invention is to provide a method for modifying the solubility of proteins by generating large numbers of genetic mutants of the gene which encodes for the protein to be solubilized which can be expressed and the resulting proteins screened for solubility.

Additional objects, advantages and novel features of the invention will be set forth in part in the description which follows, and in part will become apparent to those skilled in the art upon examination of the following or may be learned by practice of the invention. The objects and advantages of the invention may be realized and attained by means of the instrumentalities and combinations particularly pointed out in the appended claims.

SUMMARY OF THE INVENTION

To achieve the foregoing and other objects, and in accordance with the purposes of the present invention, as embodied and broadly described herein, the method for determining the solubility of a protein, P, of this invention may include the steps of: fusing a DNA fragment, [P], which codes for the protein with the DNA [R] which codes for a reporter protein, R, which can be detected in solution, forming thereby a fusion DNA fragment, [P-R], which codes for the fusion protein, P-R, such that the solubility of the P-R is determined by the solubility of protein, P; ligating the [P-R] fragment into an expression vector to form a plasmid DNA; and introducing the plasmid DNA into an expression host such that the fusion protein is overexpressed therein; whereby if the fusion protein P-R is in solution in the host, the reporter protein R can be detected, thereby indicating that the protein P is soluble.

Preferably, the DNA fragment [P] is fused with the DNA fragment [L] which codes for a flexible linker peptide, L, which has been fused with the DNA fragment [R], forming thereby either fusion DNA fragment [P-L-R] or fusion DNA fragment [R-L-P], such that the solubility of the fusion proteins encoded by the [P-L-R] or the [R-L-P] are determined by the solubility of protein P.

Preferably also, the DNA fragment bearing [L-R] or [R-L] is part of an expression vector and/or transfection/transformation vector enabling the fusion of [P] to yield the

DNA fusions [P-L-R] or [R-L-P] as part of said vectors, thus enabling a host cell to express either the fusion protein P-L-R or the fusion protein R-L-P, such that the solubility of the fusion protein is determined by the solubility of protein P.

It is also preferred that the linker peptide is short, flexible, hydrophilic and soluble.

Preferably also, the reporter protein includes green fluorescent protein.

In another aspect of the present invention, in accordance with its objects and purposes, the method for modifying the solubility of a protein, P, hereof may include the steps of: introducing mutations into [P], the DNA fragment which codes for the protein, generating thereby a combinatorial library of mutated variants, [X]; in-frame fusing individual [X] variants with a DNA construct such as a plasmid vector which includes a fragment which codes for a reporter protein, [R], which can be detected in solution, forming thereby a set of DNA constructs containing [X-R], which code for the fusion proteins, X-R, such that the solubility of each of the X-R proteins is determined by the solubility of the variant protein X contained therein; and introducing each of the DNA constructs into an expression host such that the fusion protein is overexpressed therein; whereby if one of the fusion proteins X-R is soluble in the host therefor, said reporter protein R can be detected, thereby indicating that the variant of the protein P is soluble.

Preferably, the DNA fragment [X] is fused with the DNA fragment which codes for a flexible linker peptide, [L], which has been fused with the DNA fragment [R], forming thereby either fusion DNA fragment [X-L-R] or fusion DNA fragment [R-L-X], such that the solubility of the fusion proteins expressed by the [X-L-R] or the [R-L-X] are determined by the solubility of protein X.

Preferably also, the DNA fragment bearing [L-R] or [R-L] is part of an expression vector and/or transfection/transformation vector enabling the fusion of [X] to yield the DNA fusions [X-L-R] or [R-L-X] as part of said vectors, thus enabling a host cell to

express either the fusion protein X-L-R or the fusion protein R-L-X, such that the solubility of the fusion protein is determined by the solubility of protein X.

It is preferred that the linker peptide short, flexible, hydrophilic and soluble.

Preferably also the reporter protein includes green fluorescent protein.

It is also preferred that the step of introducing mutations into [P] generating thereby a combinatorial library of mutated variants [X] is achieved using gene shuffling and directed evolution.

Benefits and advantages of the present invention include the enhancement of the solubility of proteins of interest without having to individually test, (such as by large-scale growth of each mutant in question followed by cell lysis, fractionation and sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE)), the solubility of each protein modification generated, and has general applicability.

BRIEF DESCRIPTION OF THE DRAWINGS

The accompanying drawings, which are incorporated in and form a part of the specification, illustrate the embodiments of the present invention and, together with the description, serve to explain the principles of the invention. In the drawings:

FIGURE 1 is a flow diagram illustrating the use of the solubility reporter according to the teachings of the present invention; if protein, P, is insoluble, the fusion protein, P-L-GFP, is insoluble, aggregated or bound in inclusion bodies, and is nonfluorescent, while if protein P is soluble, fusion protein P-L-GFP is soluble and fluorescent.

FIGURE 2 is a flow diagram illustrating the generation of mutated versions of an arbitrary protein, P, which have enhanced solubility, employing fluorescence-assisted cell sorting to identify and select mutants with enhanced solubility.

FIGURE 3 illustrates the performance of the GFP solubility reporter in *E. coli* BL21(DE3) induced by isopropyl-β-D-thiogalactopyranoside (IPTG) on Luria-Bertani (LB) media plates.

FIGURE 4 illustrates the increase in fluorescence of clones expressing the fusion H-type ferritin-L-GFP during the process of directed evolution using nutrient agar plates.

FIGURE 5 illustrates the application of the method to improve the solubility of bullfrog H-type ferritin, a protein which is normally insoluble when overexpressed at 37° C in *E. coli*.

DETAILED DESCRIPTION

Briefly, the present invention utilizes a solubility reporter protein, expressed by the DNA fragment [R], which gives a specific, measurable signal when the protein encoded by the in-frame fusion DNA fragment, [P-L-R], is soluble, where [P] is the DNA fragment which encodes the protein, P, to be solubilized, and [L] is the DNA fragment which encodes flexible linker peptide, L. In one embodiment of the invention, R is green fluorescent protein (GFP). Linker peptide L, which is preferably optimized for flexibility, hydrophilic nature, and solubility, is fused to the GFP. When overexpressed in the host cell, for example *E. coli*, the fusion protein(s) L-GFP (GFP fused to the C-terminus of L) or GFP-L (GFP fused to the N-terminus of L) are soluble within the expression host and fluorescent. The DNA encoding P is then fused to a reporter vector containing the DNA fragment which encodes the L-GFP construct, and the fusion protein P-L-GFP (P fused to the N-terminus of L-GFP) is caused to be overexpressed in host cell. Alternatively, the DNA encoding P is fused to a reporter vector containing a DNA fragment which encodes the GFP-L construct, and the fusion protein GFP-L-P (P fused to the C-terminus of GFP-L) is caused to be overexpressed in the host cell. The GFP-L and L-GFP are chosen such that the solubility of the P-L-GFP or GFP-L-P is controlled by the solubility of P. It is anticipated that for some systems, linker peptide L will not be required. When P is soluble, the proteins P-L-GFP or GFP-L-P are soluble within the expression host and are fluorescent. When P is insoluble, P-L-GFP or GFP-L-P are found in aggregates within the host known as inclusion bodies and are non-fluorescent. Thus, P-L-GFP or GFP-L-P constitute solubility reporters for rapidly determining the solubility of P. Figure 1 is a schematic representation of the use of the solubility reporter according to the teachings of the present invention.

Modification and, more particularly, enhancement of the solubility of protein P is accomplished by use of a DNA construct containing at least the solubility reporter DNA

fragments [L-GFP] or [GFP-L], in a directed evolution of [P]. A combinatorial library of mutated variants X is generated by gene shuffling, for example. The resulting pool of genes [X] encoding mutated proteins X is then genetically fused in-frame either with a pool of DNA constructs such as vectors containing [L-GFP] to produce a pool of DNA constructs encoding fusion proteins X-L-GFP; or to a pool of DNA constructs containing [GFP-L] to produce a pool of DNA constructs encoding fusion proteins GFP-L-X, each fusion variant having solubility determined by X. After introducing the DNA into an expression host, such as electroporation of circular plasmid vectors into *E. coli*, individual variants with increased fluorescence (and hence increased solubility) may be screened and separated using fluorescence-assisted cell sorting, as an example. Millions of variants can be screened in one hour. Further cycles of directed evolution may be instigated until no further improvement in solubility is observed. Furthermore, mutations which are unnecessary for enhanced solubility which accumulated during the directed evolution, can be removed by *in vitro* recombination or backcrossing of the DNA encoding enhanced variants X of P against an excess of DNA encoding wild type P, followed by selection of variants retaining enhanced solubility, using said solubility reporter. Figure 2 is a schematic illustration of the generation of mutated versions of an arbitrary protein, P, which have enhanced solubility, employing fluorescence-assisted cell sorting (FACS) to identify and select mutants with enhanced solubility according to the teaching of the present invention.

To screen large numbers of versions of an arbitrary protein, it is desirable, but not essential, that reporter R be chosen to have the following characteristics: (1) The observed parameter for R, which indicates solubility of X-L-R and R-L-X, must not be observable independent of the solubility of X or by the presence of X; (2) R should not dominate the solubility of X-L-R; (3) The solubility of X-L-R and R-L-X should be determined primarily by the solubility of X; (4) R should not assist the folding of X; (5) L should not significantly influence the solubility of R-L-X or X-L-R; and (6) L should not dominate the folding of any of X, R, X-L-R, or R-L-X.

Having generally described the invention, the following EXAMPLES illustrate the application of the method of the present invention in greater detail.

EXAMPLE 1

As an example of the assembly of a construct which satisfies the above-described six criteria, a Bgl-II/Xho-1 fragment of plasmid pET-21a(+), containing: the T7 promoter; lac operator sequence; ribosomal binding site; and multiple cloning site was ligated into the Bgl-II/Xho-1 site of pET-28a(+). The resulting hybrid plasmid contained the Kan, lacI, and F1 origin of replication of the pET-28a(+) backbone. The pET21a(+) and pET28a(+) vectors were used as obtained from a commercial source. The vector was digested with Nde-1 and BamH-1, the small fragment was discarded, and replaced with an in-frame stuffer such that the sequence, inclusive of the Nde-1 and BamH-I sites, was [CATATGTGTAGACAGCTGGGATCC]. Next, the vector was digested with BamH-I and EcoR-1 and the small stuffer was discarded. The BamH-I/EcoR-1 site was filled with the DNA fragment [GGATCCGCTGGCTCCGCTGCTGGTTCTGGCGAATTC], coding for the flexible linker L (GSAGSAAGSGEF). An improved variant of GFP was created by site-directed mutation using recombinant PCR (see, e.g., "Recombinant PCR" by Russel Higuchi in "PCR Protocols, a Guide to Methods and Applications", Michael A. Innis, David H. Gelfand, John J. Sninsky, and Thomas J. White, eds. Academic press, Inc., 177, (1990)), of the soluble variant of Crameri et al., *supra*, to yield the red-shift S65T mutation (See, e.g., "Improved Green Fluorescence," by Roger Heim et al., Nature **373**, 663, (1995)) which improves the performance of the protein in FACS, by increasing the absorption of the fluorophore of 488 nm light (near the argon laser emission commonly used for FACS). The internal Nde-1 and BamH-1 sites were abolished by silent-mutation. The resulting GFP variant was amplified by PCR using the 5' primer [GATATAGAATTCAGCAAAGGAGAAGAACTTTTC], incorporating a 5' EcoR-1 site; and the 3' primer [GAATTCGGTACCTTATTTGTAGAGCTCTACCAT], incorporating a 5' Xho-1 site. The resulting vector was digested with EcoR-1/Xho-1, the stuffer discarded, and replaced with the EcoR-1/Xho-1-digested EcoR-1:GFP:Xho-1 amplicon, and the circular plasmid produced thereby was transformed by

electroporation into the *E. coli* strain BL21(DE3) genotype: (F^- *ompT* $hsdS_B$ ($r_B^-m_B^-$) *gal dcm* (DE3)), a commercially available strain. The construct in the pET vector system is inducible by IPTG. A transformant was used to inoculate a culture of LB and grown to an optical density (O.D.) at 600 nm of approx. 0.5, IPTG was added to a final concentration of 1 mM, and induction was allowed to proceed for 2 h. The bright green fluorescence, visible under room lighting, indicated that the fusion construct was soluble and well-expressed. Next, the small in-frame stuffer fragment between Nde-1 and BamH-1 was removed by restriction digest, and replaced by an out-of-frame stuffer with 3 translational stops. Cells expressing this fusion were non-fluorescent due to termination of translation prior to the GFP. Finally, the vector was digested with Nde-1+BamH-1 to remove the stuffer and create a recipient site for Nde-1/BamH-1 flanked inserts. This recipient vector is subsequently referred to as the solubility-reporter vector. The specific examples described below use primers for the genes of interest which contain Nde-1(N-terminus) and BamH-1 (C-terminus). The use of an out-of-frame stuffer insures that and vectors escaping digest code for non-fluorescent constructs and thus had the effect of eliminating false-positives.

The response of the reporter system prepared as described hereinabove to two proteins (one highly soluble, the other highly insoluble) which are each efficiently overexpressed in *E. coli* is demonstrated in Fig. 3. A fusion to the highly soluble protein malE, which is widely used as a fusion protein to facilitate the purification of overexpressed proteins in *E.coli*, [malE-L-GFP], was selected to demonstrate the response of the reporter system to a soluble protein. A fusion construct with xylR, a highly insoluble bacterial regulator protein, [xylR-L-GFP], was chosen to demonstrate the response of the reporter system to an insoluble protein. The constructs were overexpressed in strain BL21(DE3), clones were allowed to grow on nitrocellulose membranes on LB media agar plates containing kanamycin until colonies were 1-2 mm in diameter. The membranes bearing the colonies were transferred to LB media agar plates containing kanamycin and the IPTG inducer to cause overexpression of the fusion proteins. Under long-wavelength UV radiation Fig. 3a is a photograph of the

resulting brightly fluorescent colonies where the protein malE-L-GFP is overexpressed, while Fig. 3b is a photograph of the resulting weakly fluorescent colonies where the protein xylR-L-GFP is overexpressed.

The response of the solubility reporter system during improvement of the solubility of bullfrog H-ferritin by directed evolution of the expressed fusion construct, [ferritin-L-GFP], is shown in Fig. 4. The 6 clones of the ninth row (from left to right) are: wild type (barely visible at the extreme left); followed by optima, (brightest, most soluble), from cycles 1, 2, 3 and 4 of directed evolution, and round 1 of backcrossing of the round 4 optima against the wildtype ferritin. The upper grid of 8 rows, 6 clones per row (48 colonies), are optima from a second round of backcrossing to remove non-essential mutations. With each cycle, the fluorescence (and hence solubility) improves.

Figure 5 shows the use of an SDS-PAGE gel to illustrate the effectiveness of solubility reporters in a directed evolution process to improve the solubility of bullfrog H-type ferritin expressed in *E. coli*. Cultures expressing non-fusion constructs of ferritin alone were sonicated to lyse cells, and the soluble and insoluble fractions were separated by centrifugation. Fractions were resolved by SDS-PAGE; here, **S** = soluble (supernatant) fraction, **P** = insoluble (pellet) fraction. Molecular weight marker ladder, **M** = 10 kDAL. Lanes **1,2** are bullfrog L-type ferritin, a soluble protein used as control; lanes **3,4** are insoluble wild-type bullfrog H-type ferritin; lanes **5,6** are the round 4 optimum variant of bullfrog H-type ferritin after 2 rounds of back-crossing against the wildtype to remove spontaneous mutations not related to solubility. Improvement of the solubility of round 4 variant is observed by comparing lane 5 with the wildtype (lane 3) H-type ferritins. Round 4 optimum (with 2 back-crossing rounds) was picked from row 3, column 2 of the plate shown in Fig. 4 hereof, and shows that the strong fluorescence from the solubility reporter is indeed related to solubility of the fusion protein construct.

EXAMPLE 2

The above-described use of a solubility reporter can be analogously extended to determine the solubility of protein fragments. For example, to determine the solubility of fragments F of a protein P, the DNA [P] is subjected to a partial enzymatic digest, (e.g.,

by DNASE-I in the presence of the divalent cations Mn^{2+} or Co^{2+}), to create a pool of smaller fragments, [F]. The fragments can be polished with a proof-reading polymerase bearing 3'-5' exonuclease activity to yield blunt-ends, or subsequently given A-overhangs by treatment with a polymerase devoid of 3'-5' exonuclease activity with excess dATP (e.g., Taq polymerase). If desired, a particular size range of the fragments [F] may be selected, by agarose gel electrophoresis as an example. After ligation (e.g., blunt-end or T/A overhang) with the pool of appropriate recipient solubility reporter vector (e.g., bearing a blunt-end or T/A cloning site in-frame with [L-R]), some of the fragments [F] will form in-frame translational fusions, [F-L-R]. After transformation into an appropriate host, (e.g., *E. coli*), expressed fusion proteins F-L-R which contain a soluble fragment F will be soluble, and detectable in the host by virtue of R (e.g., if R is GFP the host cells will be fluorescent). Thus, the above-described solubility reporter method may be used to determine the solubility of a protein, its variants (mutants), and fragments thereof.

EXAMPLE 3

EXAMPLE 1 has shown that GFP can be used as a solubility reporter. However, solubility reporters incorporating a translational fusion [P-L-R] include systems in which R is a protein/peptide other than GFP. When the fusion construct [P-L-R] is used, R can be a protein/peptide which gives a detectable signal observable by chemical, biological or physical means, when linked to P-L as P-L-R. As an example, R could be the beta-galactosidase enzyme, lacZ. Clones expressing P-L-lacZ in which P is a soluble protein are detected by the enzymatic activity of lacZ (See, e.g., "Beta-Galactosidase Gene Fusions For Analyzing Gene Expression In *Escherichia Coli* And Yeast," by M. Casadaban et al., Methods Enzymol. **100**, 293 (1983)) on substrates which yield a colored reaction product (For example, X-gal (5-bromo-4-chloro-3-indolyl-β-D-galactoside)). Colonies expressing fusion proteins with β-galactosidase activity turn blue on plates containing X-gal. Furthermore, in situations where the lacZ protein proves too large, the functionally complementable lacZα fragment is used as a

substitute. The complementary fragment Δ-lacZ is provided by the host chromosome (For example, *E. coli* strain DH10B (F⁻ *mcr*A Δ(*mrr*'*hsd*RMS-*mcr*BC) ϕ80d*lacZ*ΔM15 Δ*lac*X74 *deo*R *rec*A1 *end*A1 *ara*D139 Δ(*ara,leu*)7697 *gal*U *gal*Kλ⁻ *rps*L *nup*G), where the complementary fragment is provided by ϕ80d*lacZ*ΔM15. Fusion proteins P-L-lacZα containing a soluble protein P are soluble and contain a correctly-folded lacZα, thereby leading to complementation of the Δ-lacZ fragment and restoration of lacZ β-galactosidase activity.

EXAMPLE 4

Reporter proteins R, which have optimal activity when present in a non-fusion context may be employed for assays. The construct P-L-C-R is generated, where C is a unique protease site. For example, C could be the viral protease cleavage site for the plum pox virus NIa protease (See, e.g., M. Martin et al., "Determination of polyprotein processing sites by amino terminal sequencing of nonstructural proteins encoded by plum pox polyvirus", Virus Res. **15**, 97, (1990)), and R is the lacZα fragment, as an example. The construct P-L-C-lacZα and the viral protease (NIa) could each be expressed under the control of separately inducible promoters on separate plasmids with compatible origins of replication. For an example of the use of multiple compatible plasmids with cloning sites under independently controlled promoters, see R. Lutz and H. Berjard, "Independent and tight regulation of transcriptional units in *E. coli* via the LacR/O, the TetR/O and AraD/I_1-I_2 regulatory elements", Nucleic Acids Res., **25(6)**, 1203, (1997). The plasmids and required *E. coli* host strains are commercially available; for example, the P-L-C-lacZα construct could be expressed under the control of the tet promoter, and the NIa gene under the control of the arabinose promoter/repressor. The plasmid(s) would be transformed into the appropriate *E. coli* host (see Lutz, *supra*), and anhydrotetracycline added to the growth medium to induce expression of P-L-C-lacZα. After accumulation of the fusion protein P-L-C-lacZα, arabinose+IPTG is added to the growth medium to induce expression of the NIa protease. P-L-C-lacZα is soluble and contains a correctly-folded lacZα domain, and P-L-C-lacZα is cleaved at site C, only if P were soluble. Subsequent release of lacZα

complements the Δ-lacZ fragment and restores lacZ β-galactosidase activity, which is detected by standard colorimetric or fluorometric assays for β-galactosidase activity. As another example, R might be an antibiotic selection marker such as the β-lactamase gene (*bla*), which confers resistance to penicillin-derived antibiotics commonly used in cloning vectors. The β-lactamase gene contains a signal peptide and is translocated to the periplasm of *E. coli*. However, proper processing of the antibiotic resistance protein and translocation to the periplasm would be impeded by N-terminus fusions, although cleavage by the protease obviates this problem. The P-L-C-β-lactamase fusion protein would be soluble only if P were soluble. Concomitant induction by both anhydrotetracycline and IPTG+arabinose would provide both the fusion protein P-L-C-β-lactamase and the viral cleavage protease NIa. In cells bearing soluble variants of P, the fusion protein P-L-C-β-lactamase would be soluble and cleaved at C by virtue of the protease NIa, releasing functional β-lactamase resistance protein, thereby conferring antibiotic resistance to the antibiotic ampicillin. Conversely, in cells bearing non-soluble variants P, the fusion protein would be insoluble, the protease cleavage site C would be buried in inclusion bodies, and thereby inaccessible to cleavage by the viral protease. Furthermore, the β-lactamase protein would be buried in inclusion bodies, misfolded and non-functional. Such cells would not have resistance to the antibiotic ampicillin. It would be apparent to those having skill in the biochemical arts that selection for cells bearing soluble variants of P (and therefore having antibiotic resistance) could be accomplished by challenging mixtures of the above-mentioned cells by supplying the selective agent (e.g., the antibiotic ampicillin) in the growth medium. Moreover, it is likewise apparent to one having skill in the art that both the fusion protein P-L-C-β-lactamase and the protease NIa must be made continuously available to confer antibiotic selection throughout the life of the cell, and thus both genes must be simultaneously induced (in this example, by providing both anhydrotetracycline and IPTG/arabinose in the growth media). Cells with antibiotic resistance will survive, thereby selecting for soluble variants of P. Furthermore, additional improvement in the solubility of such variants could be accomplished by increasing the concentration of

selective agent (e.g. ampicillin) during subsequent rounds of recombination and selection.

The foregoing description of the invention has been presented for purposes of illustration and description and is not intended to be exhaustive or to limit the invention to the precise form disclosed, and obviously many modifications and variations are possible in light of the above teaching. For example, it would be apparent one having skill in biochemistry after reviewing the present disclosure that the method of the present invention can be implemented in insect, yeast and mammalian cells, wherein fusion proteins P-L-GFP are expressed to create a solubility reporter. Similarly, directed evolution for improving the solubility of proteins can be performed using insect cells, and the required DNA manipulation according to the teachings of the present invention can be achieved *in vitro* or *in vivo*.

The embodiments were chosen and described in order to best explain the principles of the invention and its practical application to thereby enable others skilled in the art to best utilize the invention in various embodiments and with various modifications as are suited to the particular use contemplated. It is intended that the scope of the invention be defined by the claims appended hereto.

WHAT IS CLAIMED IS:

1. A method for determining the solubility of a protein, P, which comprises the steps of:

(a) fusing a DNA fragment, [P], which codes for said protein with the DNA [R] which codes for a reporter protein, R, which can be detected, forming thereby a fusion DNA fragment, [P-R], which codes for the fusion protein, P-R, such that the solubility of P-R is determined by the solubility of protein, P; and

(b) introducing said DNA into an expression system such that fusion protein P-R is overexpressed therein; whereby if fusion protein P-R is soluble in said expression system, reporter protein R can be detected, thereby indicating that protein P is soluble.

2. The method for determining the solubility of a protein as described in claim 1, wherein DNA fragment [P] is fused with the DNA fragment which codes for a flexible linker peptide, [L], which has been fused with DNA fragment [R], forming thereby a fusion DNA fragment selected from the group consisting of [P-L-R] and [R-L-P], such that the solubility of fusion proteins expressed by said [P-L-R] and said [R-L-P] are determined by the solubility of said protein P.

3. The method for determining the solubility of a protein as described in claim 2, wherein linker peptide, L is chosen to be short, flexible, hydrophilic and soluble.

4. The method for determining the solubility of a protein as described in claim 1, wherein reporter protein R is selected from the group consisting of green fluorescent protein and variants thereof, lacZ, the lacZ-α fragment, and selectable marker proteins.

5. The method for determining the solubility of a protein as described in claim 4, wherein said lacZ and lacZ-α fragments include enzymes having chromogenic and fluorogenic substrates.

6. The method for determining the solubility of a protein as described in claim 4, wherein said selectable marker proteins are selected from the group consisting of ampicillin resistance proteins, tetracycline resistance proteins, kanamycin resistance proteins and arsenic resistance proteins.

7. The method for determining the solubility of a protein as described in claim 1, wherein said protein is a fragment of a larger protein and the DNA which codes for said fragment, is a fragment of the DNA which codes for said larger protein.

8. The method for determining the solubility of a protein as described in claim 7, wherein said DNA fragments which encode protein fragments of a larger protein are generated using methods from the group consisting of partial DNASE digest, radiation-induced fragmentation, chemical fragmentation, enzymatic digest, endonuclease digest, exonuclease digest, acoustic/mechanical shearing, and fragmentation.

9. The method for determining the solubility of a protein as described in claim 8, wherein said DNA fragments are size selected before said step of fusing said DNA fragment with the DNA [R] which codes for a reporter protein, R, using methods selected from the group consisting of polyacrylamide gel electrophoresis, agarose gel electrophoresis, capillary electrophoresis, and high pressure liquid chromatography.

10. A method for modifying the solubility of a protein, P, which comprises the steps of:

(a) introducing mutations into [P], the DNA fragment which codes for said protein, generating thereby a combinatorial library of mutated variants, [X];

(b) in-frame fusing individual [X] variants, with a DNA construct which contains [R] which codes for a reporter protein R which can be detected in solution, forming thereby a set of DNA constructs containing [X-R], which code for the fusion proteins, X-R, such that the solubility of each of said X-

R proteins is determined by the solubility of variant protein, X contained therein; and

(c) introducing each of said DNA constructs into an expression host such that fusion proteins X-R are overexpressed therein; whereby if one of said fusion proteins X-R is soluble in said host therefor, said reporter protein R can be detected, thereby indicating that the mutated variant of said protein P is soluble.

11. The method for modifying the solubility of a protein P as described in claim 10, wherein DNA fragment [X] is fused with the DNA fragment which codes for a flexible linker peptide, [L], which has been fused with said DNA fragment [R], forming thereby a fusion DNA fragment selected from the group consisting of [X-L-R] and [R-L-X], such that the solubility of said fusion proteins expressed by said [X-L-R] and said [R-L-X] are determined by the solubility of protein X.

12. The method for modifying the solubility of a protein as described in claim 11, wherein linker peptide, L is chosen to be short, flexible, hydrophilic and soluble.

13. The method for modifying the solubility of a protein as described in claim 10, further comprising the step of collecting said expression hosts expressing X, a more soluble form of protein P than the form of protein P expressed by the wild-type DNA.

14. The method for modifying the solubility of a protein as described in claim 13, wherein said expression hosts containing a soluble form X of protein P are separated by fluorescence assisted cell sorting from said expression hosts which contain an insoluble form X of protein P, before said step of collecting said expression hosts expressing X.

15. The method for modifying the solubility of a protein as described in claim 13, wherein said expression hosts containing a soluble form X of protein P are separated from said expression hosts which contain an insoluble form X of

protein P using nutrient agar plates, before said step of collecting said expression hosts expressing X.

16. The method for modifying the solubility of a protein as described in claim 10, wherein reporter protein R is selected from the group consisting of green fluorescent protein and variants thereof, lacZ, the lacZ-α fragment, and selectable marker proteins.

17. The method for modifying the solubility of a protein as described in claim 16, wherein said lacZ and said lacZ-α fragments include enzymes having chromogenic and fluorogenic substrates.

18. The method for modifying the solubility of a protein as described in claim 16, wherein said selectable marker proteins are selected from the group consisting of ampicillin resistance proteins, tetracycline resistance proteins, kanamycin resistance proteins and arsenic resistance proteins.

19. The method for modifying the solubility of a protein as described in claim 10, wherein said step of introducing mutations into [P], thereby generating a combinatorial library of mutated variants [X], includes methods selected from the group consisting of recombination, error-prone PCR, propagation in error-prone host strains, doping mutagenesis, saturation mutagenesis, chemical mutagenesis, irradiation mutagenesis, site-directed mutation, and combinations thereof.

20. The method for modifying the solubility of a protein as described in claim 13, further comprising the step of recombining the DNA encoding [X] from each of said collected expression hosts expressing a soluble form of said protein P, thereby yielding a pool of variant DNA fragments [X] encoding mutants X of said protein P with further enhanced solubility.

21. The method for modifying the solubility of a protein as described in claim 20, wherein said step of recombining the DNA encoding variants [X] with enhanced solubility is accomplished using recombination.

22. The method for modifying the solubility of a protein as described in claim 21, wherein the recombination is achieved by *in vitro* by gene shuffling.

23. The method for modifying the solubility of a protein as described in claim 21, wherein the recombination is achieved *in vivo* by cell-mediated recombination.

24. The method for modifying the solubility of a protein as described in claim 10, wherein mutations which do not improve solubility are removed from the DNA encoding protein X by recombination of the DNA encoding protein X with wild type DNA fragments, followed by selection for the most soluble variants.

25. The method for modifying the solubility of a protein as described in claim 10, wherein said protein is a fragment of a larger protein and the DNA which codes for said fragment, is a fragment of the DNA which codes for said larger protein.

26. The method for modifying the solubility of a protein as described in claim 25, wherein said DNA fragments which encode protein fragments of a larger protein are generated using methods from the group consisting of partial DNASE digest, radiation-induced fragmentation, chemical fragmentation, enzymatic digest, endonuclease digest, exonuclease digest, acoustic/mechanical shearing, and fragmentation.

27. The method for modifying the solubility of a protein as described in claim 26, wherein said DNA fragments are size selected before said step of fusing said DNA fragment with the DNA [R] which codes for a reporter protein, R, using methods selected from the group consisting of polyacrylamide gel electrophoresis, agarose gel electrophoresis, capillary electrophoresis, and high pressure liquid chromatography.

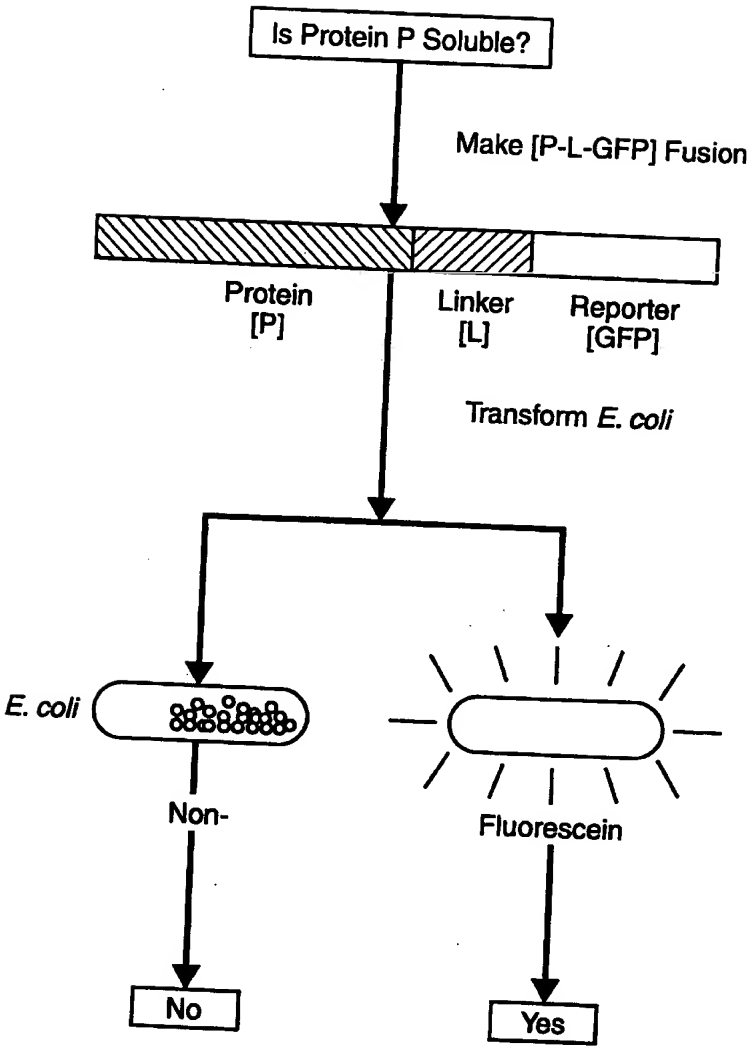
Is Protein P Soluble?
Make [P-L-GFP] Fusion
Protein [P]
Linker [L]
Reporter [GFP]
Transform E. coli
E. coli
Non-
Fluorescein
No
Yes

Fig. 1

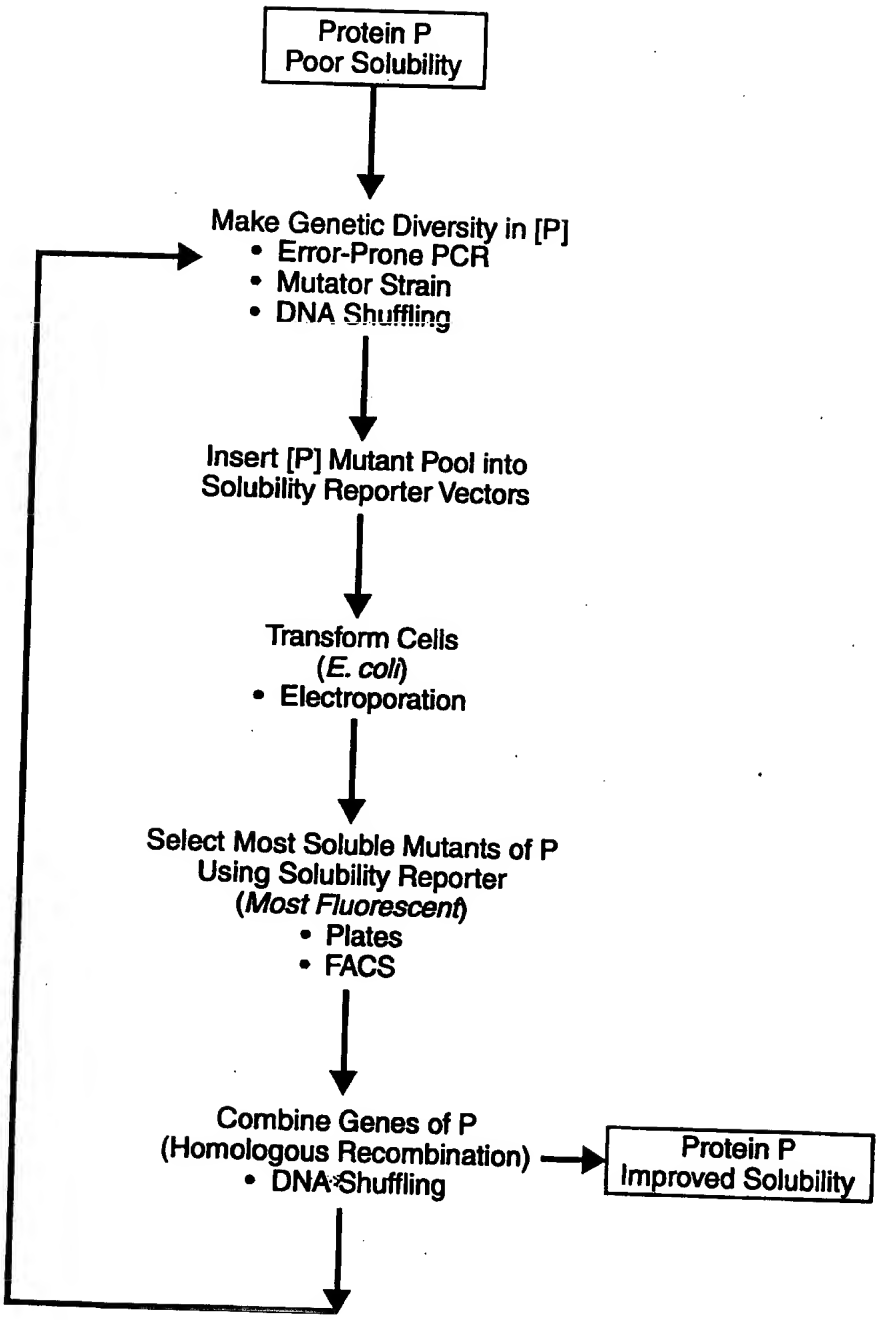
Protein P
Poor Solubility
Make Genetic Diversity in [P]
• Error-Prone PCR
• Mutator Strain
• DNA Shuffling
Insert [P] Mutant Pool into
Solubility Reporter Vectors
Transform Cells
(E. coli)
• Electroporation
Select Most Soluble Mutants of P
Using Solubility Reporter
(Most Fluorescent)
• Plates
• FACS
Combine Genes of P
(Homologous Recombination)
• DNA Shuffling
Protein P
Improved Solubility

Fig. 2

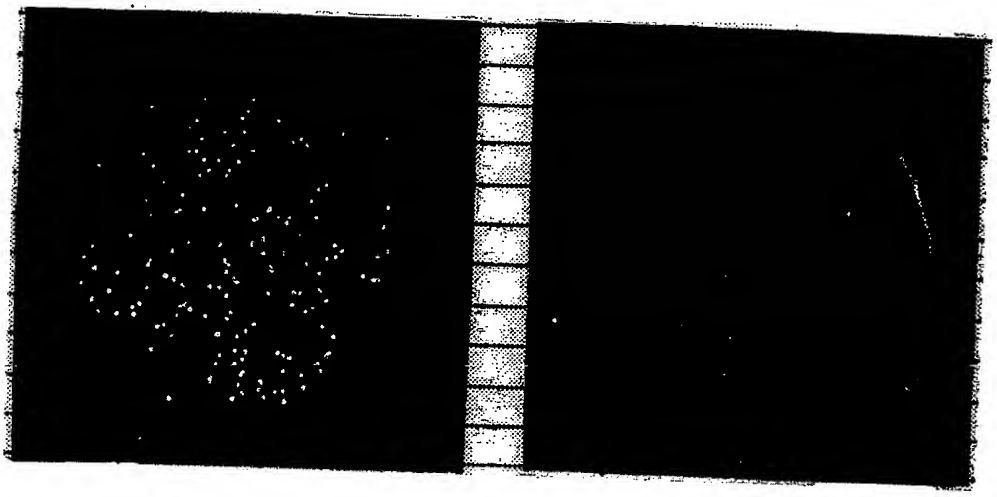

Fig. 3a

Fig. 3b

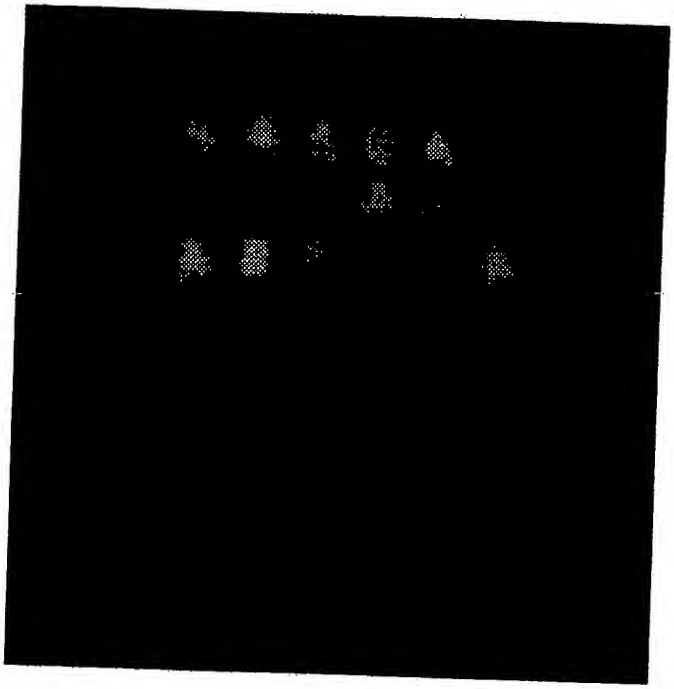

Fig. 4

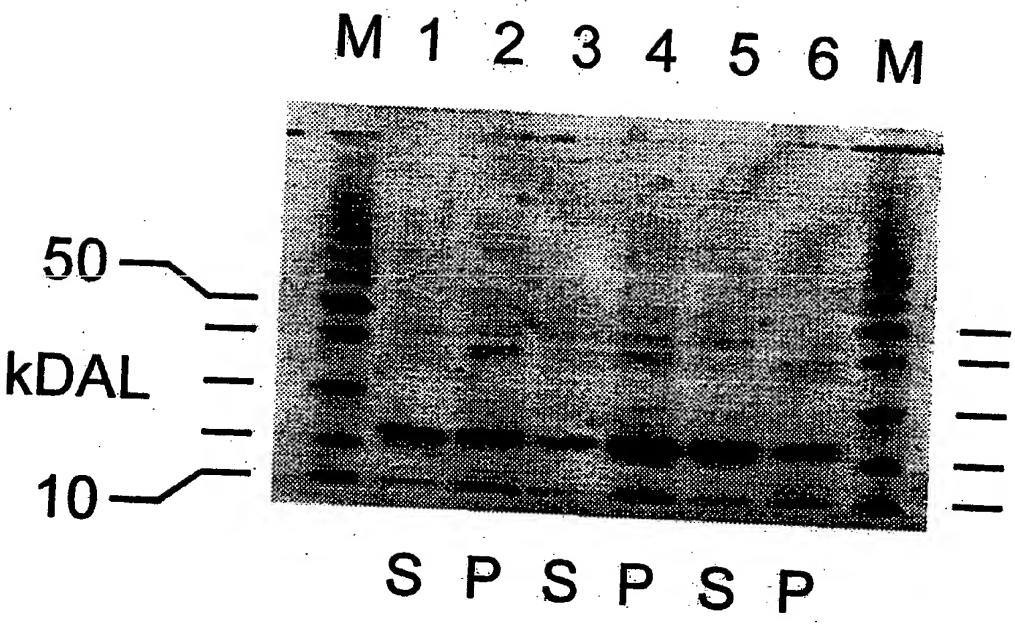

Fig. 5

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US98/25862

A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) :C12Q 1/00, 1/34, 1/02; C12P 21/06, 21/04
US CL : 435/4, 18, 29, 69.1, 69.7, 440

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 435/4, 18, 29, 69.1, 69.7, 440

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

Please See Extra Sheet.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X --- Y	US 5,491,084 A (CHALFIE et al.) 13 February 1996, col. 5, lines 44-56, col. 12, lines 28-42.	1,4,7 --- 2,3,5,6,8-27
X --- Y	WU, C. et al. Novel green fluorescent protein (GFP) baculovirus expression vectors. Gene. 29 April 1997, Vol. 190, pages 157-162, especially pages 161 and 162.	1,4,7 --- 2,3,5,6,8-27
Y	STEMMER, W. Rapid evolution of a protein in vitro by DNA shuffling. Nature. 04 August 1994, Vol. 370, pages 389-391, see entire document.	10-27
Y	CRAMERI, A. et al. Combinatorial Multiple Cassette Mutagenesis Creates all the Permutations of Mutant and Wild-Type Sequences. BioTechniques. 1995, Vol.18, No. 2, pages 194-196, see entire document.	10-27

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
29 MARCH 1999	28 APR 1999
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703) 305-3230	Authorized officer PETER TUNG Telephone No. (703) 308-0196

Form PCT/ISA/210 (second sheet)(July 1992)*

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US98/25862

B. FIELDS SEARCHED

Electronic data bases consulted (Name of data base and where practicable terms used):

APS, MEDLINE, CANCERLIT, CAPLUS, BIOSIS, EMBASE, WPIDS

search terms: protein solubility, green fluorescent protein, reporter, galactosidase, fusion protein, combinatorial library, in vitro mutagenesis, inclusion body, Geoffrey Waldo